Mining Association Rules in Data Warehouses

Haorianto Cokrowijoyo Tjioe, Monash University, Australia
David Taniar, Monash University, Australia

ABSTRACT

Data mining applications have enormously altered the strategic decision-making processes of organizations. The application of association rules algorithms is one of the well-known data mining techniques that have been developed to cope with multidimensional databases. However, most of these algorithms focus on multidimensional data models for transactional data. As data warehouses can be presented using a multidimensional model, in this paper we provide another perspective to mine association rules in data warehouses by focusing on a measurement of summarized data. We propose four algorithms — $V\text{Avg}_{\text{g}}$, $H\text{Avg}_{\text{g}}$, $W\text{MAvg}_{\text{g}}$, and ModusFilter — to provide efficient data initialization for mining association rules in data warehouses by concentrating on the measurement of aggregate data. Then we apply those algorithms both on a non-repeatable predicate, which is known as mining normal association rules, using GenNLI, and a repeatable predicate using ComDims and GenHLI, which is known as mining hybrid association rules.

Keywords: association rules; data mining; data preprocessing; data warehouse; multidimensional database design

INTRODUCTION

Association rules on transaction database were first introduced by Agrawal (1993). By using its Apriori algorithm, large itemsets satisfying the minimum support and association rules based on the minimum confidence could be discovered. Since then, a large number of efficient algorithms using the hash-based technique (Park, Chen, & Yu, 1995), transaction reduction (Han & Fu, 1995), the partition technique (Mannila, Toivonen, & Verkamo, 1994), and sample datasets to prune the number of passes on the data (Toivonen, 1996) have been introduced.

Association rules traditionally use transactional data that focus on a single dimension or predicate (Agrawal & Srikant, 1994; Han & Fu, 1995; Mannila, Toivonen, & Verkamo, 1994; Park, Chen, & Yu, 1995; Savasere, Omiecinski, & Navathe, 1995). However, this is not adequate since real life data usually involves more than one dimension or predicate. Subsequently, traditional association rules were developed to
solve the multidimensional model (Guenzel, Albrecht, & Lehner, 1999; Kamber et al., 1997). Kamber et al. (1997) exposed the idea of mining association rules in a multidimensional data model. Their algorithm focuses only on presenting association rules in a multidimensional model, which involves more than one dimension in transactional data. However, this algorithm did not discuss the hierarchies that are also characteristic of a multidimensional model. Later on, a new algorithm (Guenzel, Albrecht, & Lehner, 1999) was proposed to support mining multidimensional databases by hierarchy using an online analytical processing (OLAP) approach.

Apparently, both concepts in Guenzel, Albrecht, and Lehner (1999) and Kamber et al. (1997) miss the most important attribute, which is the measurement of aggregate data in a Data Warehouse (DW). The data in a DW contains only summarised data such as quantity sold, amount sold, and etcetera. No transactions data is stored. In this paper, we focus on providing a framework for mining association rules both on non-repeatable predicates and repeatable predicates in data warehouses by concentrating on the measurement of aggregate data.

Here, we propose four algorithms — HAvg, VAvg, ModusFilter, and WMAvg — to provide efficient data initialisation for mining association rules in data warehouses by concentrating on the measurement of aggregate data, specifically its quantity. Then we apply those algorithms both to non-repeatable predicates using GenNLI, and repeatable predicates using ComDims, and GenHLI, which is known as mining hybrid association rules.

As shown in Figure 1, we provide a framework for mining association rules in a data warehouse. We use quantity data in a fact table to explain our approach. There are three steps to perform in order to derive an initialised data for mining association rules in a data warehouse. First, we select the data warehouse that we want to mine. Secondly, we use user input variables to decide the dimensions that will be used along with the single or interval data to find the interesting patterns. Finally, we use our approach: HAvg, VAvg, ModusFilter, and WMAvg algorithms to produce data initialisation based on the information derived from the two previous steps (see Figure 1). Then we mine the DW using data initialisation from those proposed algorithms to mine non-repeatable association rules and the repeatable predicate, which is known as mining hybrid association rules.

HAvg, VAvg and WMAvg algorithms work by selecting the average measurement of aggregate data in a DW with multidimensional structures such as average quantity sold, average price, and so forth. We prune all the rows in the fact table, which have less than the average quantity, since we assume that rows with quantities less than its average will not form any association rules. The main differences between these algorithms are: HAvg finds the average quantity of the defined dimensions horizontally while VAvg finds the average quantity of the defined dimensions vertically and WMAvg selects the weighted moving average quantity of the defined dimensions vertically.

Using the VAvg algorithm, we find the average quantity of the selected dimension vertically. We illustrate how the VAvg algorithm works in Table 1. Here we use only two dimensions: Times dimension for a week’s sales only, and Products dimension for only Beer and Bread. Times dimension works as a grouping dimension. So the attributes of Products dimension will be grouped according to its time frames (see Figure 2 for the detail of those dimensions).
33 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the product's webpage:

www.igi-global.com/article/mining-association-rules-data-warehouses/1755?camid=4v1


www.igi-global.com/e-resources/library-recommendation/?id=2

Related Content

Image Mining: Detecting Deforestation Patterns Through Satellites
Marcelino Pereira dos Santos Silva, Gilberto Câmara and Maria Isabel Sobral Escada (2009). Data Mining Applications for Empowering Knowledge Societies (pp. 55-75).
www.igi-global.com/chapter/image-mining-detecting-deforestation-patterns/7546?camid=4v1a

Recent Developments on Security and Reliability in Large-Scale Data Processing with MapReduce
Christian Esposito and Massimo Ficco (2016). International Journal of Data Warehousing and Mining (pp. 49-68).
www.igi-global.com/article/recent-developments-on-security-and-reliability-in-large-scale-data-processing-with-mapreduce/143715?camid=4v1a
A Presentation Model & Non-Traditional Visualization for OLAP
Andreas Maniatis, Panos Vassiliadis, Spiros Skiadopoulos, Yannis Vassiliou, George
Warehousing and Mining (pp. 1-36).
www.igi-global.com/article/presentation-model-non-traditional-
visualization/1746?camid=4v1a

PAKDD-2007: A Near-Linear Model for the Cross-Selling Problem
Thierry Van de Merckt and Jean-François Chevalier (2010). Strategic Advancements
in Utilizing Data Mining and Warehousing Technologies: New Concepts and
Developments (pp. 320-328).
www.igi-global.com/chapter/pakdd-2007-near-linear-
model/40414?camid=4v1a